

Predicating depression on Twitter using hybrid model BiLSTM-XGBOOST

Rula Kamil¹, Ayad R. Abbas²

¹Department of Software, College of Information Technology, University of Babylon, Al-Hillah, Iraq

²Department of Applied Computing, Faculty of Computer Science and Environment, University of Technology, Baghdad, Iraq

Article Info

Article history:

Received Dec 4, 2022

Revised Feb 23, 2023

Accepted Mar 7, 2023

Keywords:

Bidirectional-long short-term
memory

Depression

Hybrid model

Twitter

XGBOOST

ABSTRACT

Nowadays, depression is a common mental illness. Failure to recognize depression early or guarantee that a depressed individual receives prompt counseling can lead to serious issues. Social media allow us to monitor people's thoughts, daily activities, and emotions, including persons with mental illnesses. This study suggested novel hybrid models that combine one of the deep learning techniques with one of the machine learning approaches. This paper used a dataset from the Kaggle website to predict depression. Two deep learning techniques were chosen to conduct the experiments: bidirectional long short-term memory (Bi-LSTM), and convolutional neural network (CNN). Three machine learning techniques were also selected, which are support vector machine (SVM), light gradient boosting machine (LGBM), and extreme gradient boosting (XGBOOST). Deep learning methods were applied to extract important features from input data and training, and then machine learning was utilized to predict the class. The performance of the hybrid models was compared against that of five single models. The results showed that Bi-LSTM-XGBOOST is better than single models and achieve the highest performance, with 94% for all evaluation metrics. The proposed model can improve the performance of machine learning techniques and increase the detection rate of depression.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Rula Kamil

Department of Software, College of Information Technology, University of Babylon

Al-Hillah, Iraq

Email: rola.kamil@uobabylon.edu.iq

1. INTRODUCTION

Nowadays, depression is a widespread illness. There are many people worldwide who suffer from depression. Around 264 million people of all ages suffer from depression, as indicated by the World Health Organization (WHO) [1]. Depression is one of the most common causes of suicide, with over 800,000 suicide deaths occurring every year; moreover, it is the second leading cause of death among individuals in the 15 to 29 years-old range [2].

Depression is a popular mental disorder [2]. Depressive disorders come in a variety of types. For each type, there are a set of symptoms. Major depressive disease (MDD) is a popular type of depressive disorder. With this type, people cannot sleep, eat, or work. During at least two weeks, each day, the patient must have at least five of the symptoms. The symptoms include a sad mood for most of the day, a lack of interest in all activities, losing or gaining weight, a lack of energy, body agitation or retardation, a feeling of remorse or worthlessness, an inability to sleep or sleep for longer periods, and thoughts of death and suicide [3]. There are many methods used in medical fields to diagnose depression, such as surveys and interviews, but they are not accurate. Despite international healthcare programs and treatment availability, the

detection rate is low, and most depressed people deny looking for treatment. Recently, social media have been utilized by people too, especially, among the younger [3]. In 2015, there were about 2 billion social media users, and that number is growing every day [4]. Users can use social networks sites (SNS) from anywhere and at any time, whether through their mobile device or computer. The availability of social media enabled users to share feelings, interests, and daily lives. Nowadays people can share their negative feeling on social media without fear [3].

Social media has become a data source in different contexts, especially in the medical field, to monitor people's health, such as mental health. It is possible to identify common signs of depression using user-generated content on social media, and this represents a new form for the screening of mental disorders [1]. Many studies mention that language patterns may be indicators of mental state and are used in the early detection of depression [1], [3], [5]. Thus, previous studies used social media in the early detection of depression. One of the most important social media sites is Twitter, which has more than 330 million active users globally [6]. This paper used Twitter to detect depression through text.

Many research used traditional machine learning techniques to detect depression in social media. However, there are some limitations, the first limitation is using a single traditional machine learning technique and obtaining low accuracy. For example, in [7] two traditional machine learning were applied, which are support vector machine (SVM) and random forest. The best accuracy obtained was 77% with random forest. According to Alsagri and Ykhlef [8], SVM, naive bayes, and decision tree were performed as classifiers. The best accuracy was 82% with SVM. Research by Kumar *et al.* [9] applied gradient boosting, multinomial naive bayes, random forest, and ensemble vote as classifiers. The outcomes show that the best accuracy was 85.09% with the ensemble vote classifier. The effectiveness of the traditional technique was limited as the volume of data grew and the number of correlations considerably increased [10].

Nowadays, deep learning techniques are commonly used in the detection of depression field. The benefit of deep learning is the ability to extract features throughout the learning process on huge datasets [11]. Many studies showed that the results of applied deep learning are better than machine learning [12]. However, each technique has benefits and drawbacks. Long short-term memory (LSTM) produces better results but it takes more time than convolutional neural network (CNN). LSTM is more accurate with long sentences [11]. Hybrid models that combine deep and machine learning techniques become widely used in the field of images classification such as CNN-extreme gradient boosting (XGBOOST) and achieved the best results. Nowadays, some research suggested a hybrid model for sentiment analysis, as in the study [11] that proposed a hybrid model CNN-LSTM-SVM for sentiment analysis. The results of the study show that CNN-LSTM-SVM outperforms other single methods and can be able to enhance the performance.

This paper attempts to create a hybrid model that combines the architecture of deep learning techniques such as bidirectional long short-term memory (Bi-LSTM) or CNN and traditional machine learning such as XGBOOST, SVM, and light gradient boosting machine (LGBM). The study attempts to exploit the ability of deep learning techniques to extract important features from texts and benefit from them in training machine learning techniques to improve accuracy. The goal behind this is to see if the accuracy will improve or not. Thus, the main contributions of the study are summarized as follows: i) new hybrid model was suggested that mix the architecture of traditional machine with deep learning algorithms. This has not been suggested in previous research; ii) achieving good prediction accuracy of more than 90% using deep learning and hybrid models; iii) comparison between the performance of hybrid models and other techniques; and iv) improving the performance of traditional machine learning techniques by combining them with deep learning techniques. This paper is organized as follows: section 2 describes the method for the proposed system, section 3 explains the results and discusses them, and the last section 4 includes a conclusion.

2. METHOD

This paper proposed a system to predicate if the tweet is depressed or not. The method for this system consists of five steps, as shown in Figure 1. It includes dataset, pre-processing, features extraction, classifier model, and evaluation.

2.1. Dataset

A dataset from the Kaggle website is used. A dataset containing more than 7,000 tweets scraped from Twitter that belongs to depressed and not depressed users [13]. Additional depressed tweets were added to improve accuracy and increase the number of depressed tweets [14]. Data details is shown in Table 1.

2.2. Pre-processing

This step includes data preparation processes before classification by models as shown in Figure 1. Pre-processing involves a set of steps. The first step is to divide the tweet into small parts called "tokens" using the tokenization technique. Second, replace emojis and emoticons with proper text instead of removing

them because they express the feelings of people. Third, removing special characters. Fourth, replace slang words with common words to have all words in one form [15]. Fifth, removing numbers and keeping only alphabet letters. Sixth, removing HTML tags, and URL links. Seventh, converting uppercase letters to lowercase. With traditional machine learning, normalization was applied, such as removing stop words. Stop words are popular terms that appear more frequently in tweets, and they are not important. There are some important stop words, such as the person pronoun (I, my, mine, myself, we, and you) [8]. Those words are widely used by depressed people to express themselves and must not be removed from tweets. In addition, negation words like "no" and "not" are not deleted from tweets. These words are more commonly used by depressed people, and they express negative feelings, for instance, "I'm not happy." This step is known as the customized removing stop words. Stemming is the process of reducing the number of words in tweets. Stemming is the process of replacing two words with the same meaning or a common root with a single word.

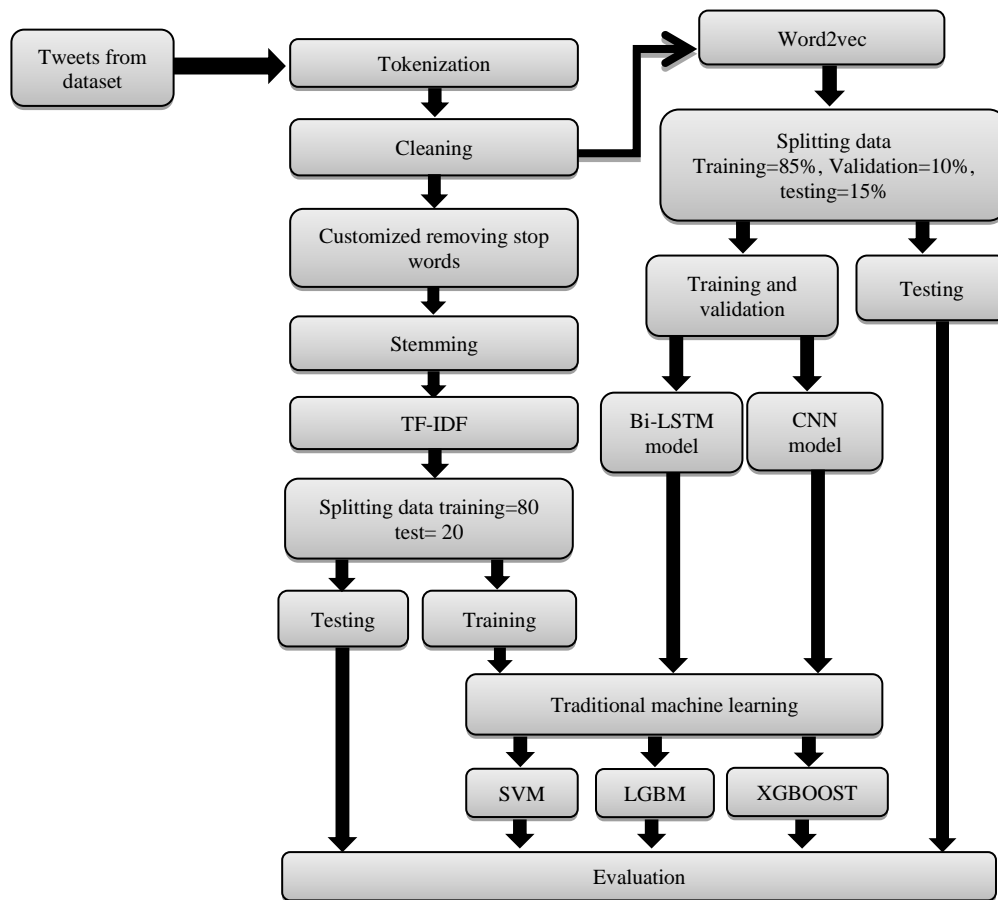


Figure 1. Method of research

Table 1. Details of the dataset

	Data1	Data2	Total tweets
Depressed tweets	3,259 [13]	2,313 [14]	5,572
Non-depressed tweets	4,694 [13]	-	4,694
			10,266

2.3. Feature extraction

To classify tweets as depressed or not by models, the text should be transformed into a numeric form that is understood by the classifier. So, it must extract features from tweets. This paper suggests using term frequency-inverse document frequency (TF-IDF) and word2vec as feature vectors.

2.3.1. Term frequency-inverse document frequency

Is a method that is commonly used to determine the importance of a word in a document. The term frequency (TF) of a particular term (t) is computed as the number of times a term (t) occurs in document (d) divided by the number of words in the document. Inverse document frequency (IDF) is utilized to identify a term's importance [16]. In (1) [16] shows how to compute IDF.

$$IDF(t) = \log [N/DF] \quad (1)$$

Where N is the number of documents and DF is the number of documents that included the term (t). TF-IDF is calculated in (2) [16]:

$$TF - IDF(t) = TF(t) \times IDF(t) \quad (2)$$

TF is relatively similar to the bag of words (BOW) method. The text from the data is represented by TF as a matrix, where rows are the total number of documents and columns are the total number of distinct words used in all of the documents. BOW ignores the order of words in documents and grammar [17]. Words with a large TF-IDF value are more significant than terms with a small TF-IDF value [17]. TF-IDF contributes to the identification of depression [8]. This method was used by [8], [18], [19].

2.3.2. Word2vec

Word2vec is one of the most well-known methods for word embedding proposed by Google in 2013. Word2vec is a neural network model. There are two learning models: a continuous bag of words (CBOW) and skip-gram. CBOW predicts the term from the context, while skip-gram predicts the context from the term [20]. Given a large corpus of text as input, word2vec produces a vector space with typically several hundred dimensions and allots a corresponding vector in the space to each unique word in the corpus. In the vector space, word vectors are located so that words with similar semantic and syntactic properties are near one another in the space. While less similar terms are placed apart from one another [21]. Word2vec's weight is determined by word order or location rather than the frequency in the same context. The similarity between the two words can be assessed after the weight of word2vec has been determined. Word2vec can represent each word as a low dimensional vector, which makes adding new words to the vocabulary list simple and easy to incorporate into new sentences [17]. About 100 billion words from the Google news dataset were used to train pre-trained vectors. For 3 million words and phrases, the model has 300-dimensional vectors. This paper uses pre-trained vectors from Google news.

2.4. Classifier model

2.4.1. Bidirectional-long short-term memory

Bi-LSTM is a type of recent neural network (RNN). Before talking about Bi-LSTM, it is necessary to know what RNN is. RNN is a type of neural network. The sample RNN includes three main layers: input, hidden layers, and output. In contrast to feedforward neural networks, the output from the past state is fed to the current state. This is useful when dealing with sequence data, such as text, and requiring remembering past words. RNNs can remember the previous words by using hidden layers, but for a small interval of time. LSTM was presented to solve this problem because it has a memory block instead of a simple RNN unit. However, LSTM can keep just the past information. Bi-LSTM was suggested to solve this issue since Bi-LSTM can keep the context for previous and future related information. So, this paper used Bi-LSTM, which can handle data in two directions: previous and next, because it works with two hidden layers [22].

2.4.2. Convolutional neural network

Is referred to as an artificial neural network. It is used to analyze and recognize images. It is specifically made for processing pixel data. There are two parts to the CNN layer: feature extraction and classification. In feature extraction, convolution series, and pooling processes are performed. In classification, a fully connected person will complete their work. It applies a filter to the input data to create a feature map. After padding the feature map, the convolutional layer is created. Max pooling is performed as a pooling layer to reduce the size of the feature map by selecting the largest value from each window. The fully connected layer is used to transform data from 2D or 3D to 1D. The fully connected layer is the last layer and is used to show the output of the network [15].

2.4.3. Light gradient boosting machine

One of the most common gradient boosting algorithms is based on a decision tree. The tree grows vertically by using a leaf-wise algorithm [23]. It was proposed by Microsoft in 2017. It was characterized by the speed of training, the need for less memory, and compatibility with big datasets.

2.4.4. XGBOOST

XGB is an ensemble tree-based technique that applied a gradient-boosting machine learning framework to solve classification and regression issues. The level-wise methods are used by XGB to grow trees. There are differences between XGB and random forest in grow of the tree, orders, and combining the results [23]. Different algorithms are used by XGB to find splits such as exact greedy and approximate algorithms that are presented first then histogram-based algorithm appeared after the LGBM method was developed. The loss value is used to determine if a split is happening or not. If the loss value exceeds a specific threshold value, then the split happens else it will ignore. This is one of the advantages leaf-wise in minimizing the number of splits with keeping the quality of a split [23].

2.4.5. Support vector machine

SVM is a supervised machine learning, non-probabilistic linear binary classifier. It generates a hyperplane in a space with high dimensions feature to split data into two classes. It finds the hyperplane that splits data into two classes with a maximum margin [24].

2.4.6. Hybrid models

This paper suggests a hybrid model that combines one of the deep learning techniques and one of the machine learning techniques. The proposed model is divided into two parts. The first part is Bi-LSTM or CNN which extracts features and information from input text. The second part receives the output of the last layer from Bi-LSTM or CNN and classifies it. Three classifiers suggested SVM, LGBM, and XGBOOST. Figures 2 and 3 explains the details of models.

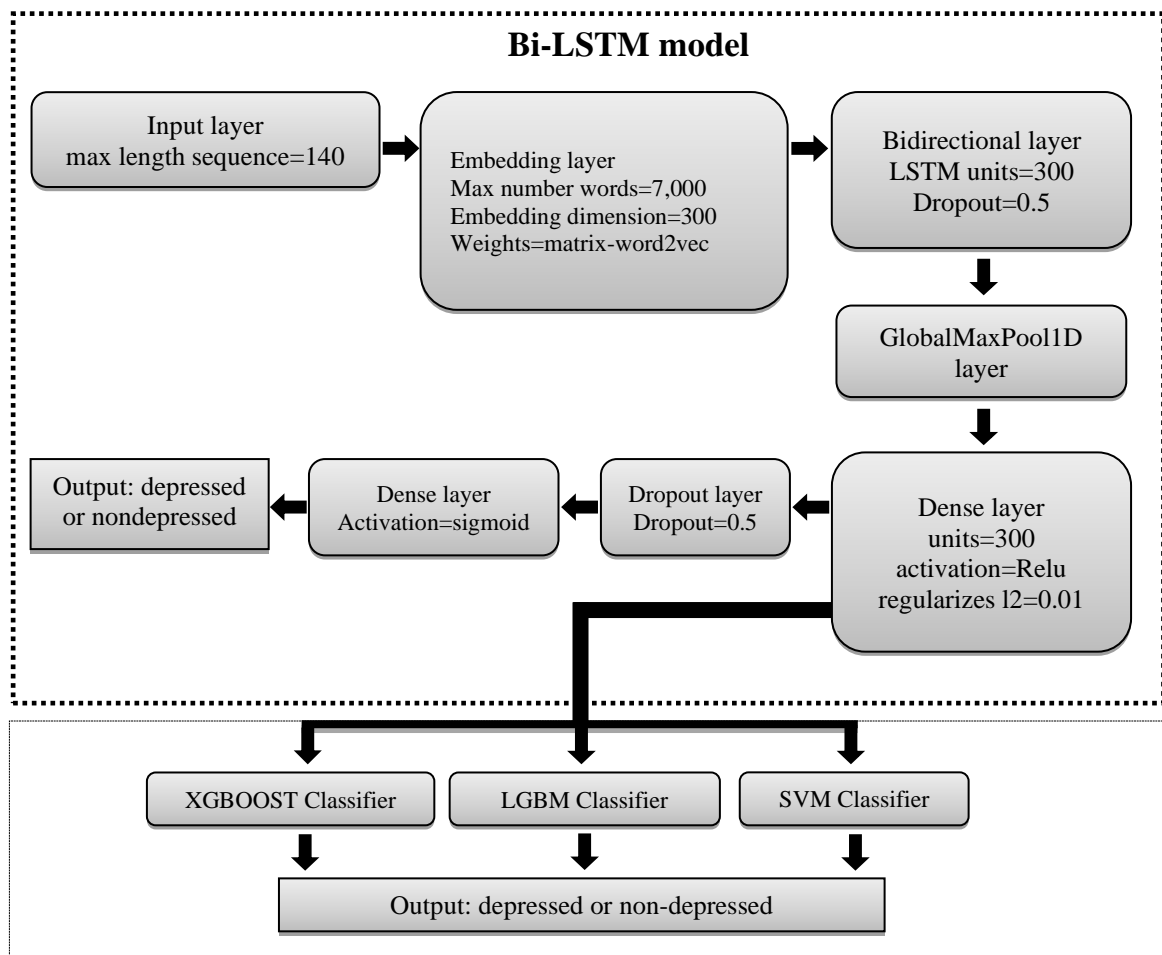


Figure 2. The hybrid Bi-LSTM traditional machine learning model

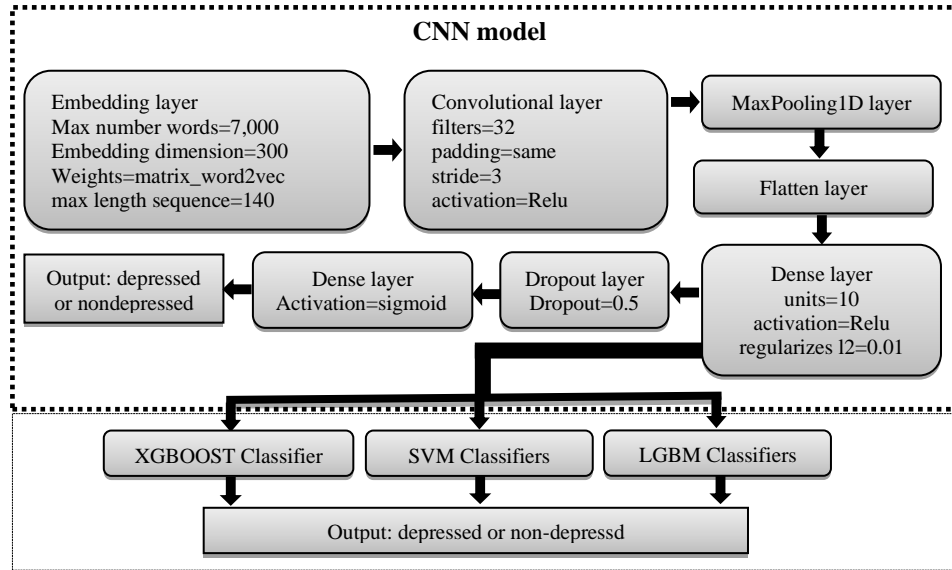


Figure 3. The hybrid CNN-traditional machine learning model

2.5. Evaluation

This research used four metrics to evaluate the performance of the model. Accuracy is the number of samples that are predicated correctly by the model divided by the total number of predictions generated by the model. It is calculated in (3):

$$Accuracy = (Tp + Tn) / (Tp + Fp + Fn + Tn) \quad (3)$$

Precision represents the number of tweets that are classified as depressed and are truly predicted to be depressed. It is computed in (4):

$$Precision = Tp / (Tp + Fp) \quad (4)$$

Recall is the number of depressed tweets that are exactly predicated by the model out of all the positive examples. In (5) explains how to find it:

$$Recall = Tp / (Tp + Fn) \quad (5)$$

F1-measure is explicated as a symmetric average of the precision and recall. In (6) shows how it was computed [25]:

$$F1 - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

3. RESULTS AND DISCUSSION

In this research, two experiments were performed. The first experiment applied traditional machine learning techniques such as SVM, LGBM, and XGBOOST with TF-IDF. The second experiment used deep learning models, hybrid models with word2vec as explained in Table 2. Four metrics were used to test and evaluate the model's performance (accuracy, precision, recall, and F1). In general, the results in Table 2 showed that hybrid models give results better than single traditional machine learning methods. SVM, LGBM, and XGBOOST with TF-IDF achieved low performance as can see in Table 2.

Furthermore, the performance of SVM, LGBM, and XGBOOST was improved when combine with CNN and Bi-LSTM, where they trained on the features extracted by CNN or Bi-LSTM. For example, the accuracy for XGBOOST is 89% when used alone, but when combined with CNN or Bi-LSTM leads to an accuracy of 93% and 94%. Also, the models that contain Bi-LSTM outperform CNN models because Bi-LSTM is effective to deal with long sentence sequences and has the memory to remember previous and future words in two directions [26].

In the last, Bi-LSTM-XGBOOST outperformed all other models with an accuracy of 0.9487. The findings for this paper match the results for [11] which show that the hybrid model CNN-LSTM-SVM outperformed other single models. In addition, extracting features from text using word2vec plays an important role. In word2vec the context of the sentence and its meaning are important.

Table 2. Results of applying traditional machine learning and hybrid model

Features extraction	Classifier	Accuracy	Precision	Recall	F1
TF-IDF	SVM	0.9303	0.9293	0.9316	0.93
	LGBM	0.9103	0.9093	0.9108	0.9099
	XGBOOST	0.8991	0.8981	0.9001	0.8987
	Bi-LSTM	0.9454	0.9497	0.9405	0.9441
	Bi-LSTM-SVM	0.9428	0.9422	0.9416	0.9419
Word2vec	Bi-LSTM-LGBM	0.9435	0.9427	0.9425	0.9426
	Bi-LSTM-XGBOOST	0.9487	0.9490	0.9466	0.9477
	CNN	0.93232	0.93071	0.93477	0.93190
	CNN-SVM	0.93928	0.93785	0.93981	0.93870
	CNN-LGBM	0.93801	0.93663	0.93839	0.93741
	CNN-XGBOOST	0.9380	0.9367	0.9382	0.9374

4. CONCLUSION

Nowadays, depression is a silent killer that eventually leads people to commit suicide. Therefore, it is necessary to bring it to a halt in any possible way. This study suggested a model that attempted to predict depression symptoms in user tweets via using hybrid machine learning and deep learning techniques. The experiments were performed with suggested models and evaluated their performance on the dataset. The outcomes are promising. The proposed hybrid models produced better outcomes than using single traditional machine learning methods and Bi-LSTM-XGBOOST outperformed other techniques. The features extracted from Bi-LSTM play a significant role in enhancing machine learning performance in addition to the role of word2vec. However, there are several limitations, such as utilizing a small dataset and using a dataset from Twitter only. In the future, application of the model to a dataset with more data and use of other types from social media such as Facebook, and different pre-trained embedding methods such as BERT and Glove.




REFERENCES

- [1] R. Safa, P. Bayat, and L. Moghtader, "Automatic detection of depression symptoms in twitter using multimodal analysis," *Journal of Supercomputing*, vol. 78, no. 4, pp. 4709–4744, 2022, doi: 10.1007/s11227-021-04040-8.
- [2] M. M. Aldarwish and H. F. Ahmad, "Predicting depression levels using social media posts," in *2017 IEEE 13th International Symposium on Autonomous Decentralized System (ISADS)*, 2017, pp. 277–280, doi: 10.1109/ISADS.2017.41.
- [3] R. M. -Castaño, J. C. Pichel, and D. E. Losada, "A big data platform for real time analysis of signs of depression in social media," *International Journal of Environmental Research and Public Health*, vol. 17, no. 13, pp. 1–23, 2020, doi: 10.3390/ijerph17134752.
- [4] M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in *Proceedings of the 5th Annual ACM Web Science Conference*, 2013, pp. 47–56, doi: 10.1145/2464464.2464480.
- [5] K. Loveys, P. Crutchley, E. Wyatt, and G. Coppersmith, "Small but mighty: affective micropatterns for quantifying mental health from social media language," in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, 2017, pp. 85–95, doi: 10.18653/v1/W17-3110.
- [6] A. Leis, F. Ronzano, M. A. Mayer, L. I. Furlong, and F. Sanz, "Detecting signs of depression in tweets in Spanish: behavioral and linguistic analysis," *Journal of medical Internet research*, vol. 21, no. 6, p. 14199, 2019.
- [7] F. Azam, M. Agro, M. Sami, M. H. Abro, and A. Dewani, "Identifying depression among twitter users using sentiment analysis," in *2021 International Conference on Artificial Intelligence (ICAI)*, 2021, pp. 44–49, doi: 10.1109/ICAI52203.2021.9445271.
- [8] H. S. Alsagri and M. Ykhlef, "Machine learning-based approach for depression detection in twitter using content and activity features," *IEICE Transactions on Information and Systems*, vol. E103D, no. 8, pp. 1825–1832, 2020, doi: 10.1587/transinf.2020EDP7023.
- [9] A. Kumar, A. Sharma, and A. Arora, "Anxious depression prediction in real-time social data," *SSRN Electronic Journal*, pp. 1–7, 2019, doi: 10.2139/ssrn.3383359.
- [10] H. Kour and M. K. Gupta, "An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM," *Multimedia Tools and Applications*, vol. 81, no. 17, pp. 23649–23685, 2022, doi: 10.1007/s11042-022-12648-y.
- [11] C. N. Dang, M. N. M. -García, and F. D. I. Prieta, "Hybrid deep learning models for sentiment analysis," *Complexity*, vol. 2021, pp. 1–16, 2021, doi: 10.1155/2021/9986920.
- [12] C. Bhargava, S. Poornima, S. Mahur, and M. Pushpalatha, "Depression detection using sentiment analysis of tweets," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 11, pp. 5411–5418, 2021.
- [13] H. K. Cho, "Twitter depression dataset," *Kaggle*, 2021. <https://www.kaggle.com/datasets/hyunkic/twitter-depression-dataset>
- [14] D. Yoo, "Detect early depression through tweets," *Kaggle*, 2022. https://www.kaggle.com/code/dongphilyoo/detect-early-depression-through-tweets/data?select=depressive_tweets_processed.csv
- [15] T. F. Mumu, I. J. Munni, and A. K. Das, "Depressed people detection from Bangla social media status using LSTM and CNN approach," *Journal of Engineering Advancements*, vol. 2, no. 1, pp. 41–47, 2021, doi: 10.38032/jea.2021.01.006.




- [16] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," *Procedia Computer Science*, vol. 152, pp. 341–348, 2019, doi: 10.1016/j.procs.2019.05.008.
- [17] P. Hoyeon and K. K. -jae, "Impact of word embedding methods on performance of sentiment analysis with machine learning techniques," *Journal of The Korea Society of Computer and Information*, vol. 25, no. 8, pp. 181–188, 2020.
- [18] K. Sudha, S. Sreemathi, B. Nathiya, and D. RahiniPriya, "Depression detection using machine learning," in *International Journal of Research and Advanced Development. on AICTE Sponsored International Conference on Data Science & Big Data Analytics for Sustainability*, 2020.
- [19] N. P. Shetty, B. Muniyal, A. Anand, S. Kumar, and S. Prabhu, "Predicting depression using deep learning and ensemble algorithms on raw twitter data," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 4, pp. 3751–3756, 2020, doi: 10.11591/ijece.v10i4.pp3751-3756.
- [20] L. Ma and Y. Zhang, "Using word2Vec to process big text data," in *2015 IEEE International Conference on Big Data (Big Data)*, 2015, pp. 2895–2897, doi: 10.1109/BigData.2015.7364114.
- [21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *Arxiv-Computer Science*, vol. 1, pp. 1–12, 2013.
- [22] H. Ahmad, M. Z. Asghar, F. M. Alotaibi, and I. A. Hameed, "Applying deep learning technique for depression classification in social media text," *Journal of Medical Imaging and Health Informatics*, vol. 10, no. 10, pp. 2446–2451, 2020, doi: 10.1166/jmihi.2020.3169.
- [23] F. Alzamzami, M. Hoda, and A. E. Saddik, "Light gradient boosting machine for general sentiment classification on short texts: a comparative evaluation," *IEEE Access*, vol. 8, pp. 101840–101858, 2020, doi: 10.1109/ACCESS.2020.2997330.
- [24] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq, "Depression detection from social network data using machine learning techniques," *Health Information Science and Systems*, vol. 6, no. 8, pp. 1–12, 2018.
- [25] A. Amanat *et al.*, "Deep learning for depression detection from textual data," *Electronics*, vol. 11, no. 5, pp. 1–13, 2022, doi: 10.3390/electronics11050676.
- [26] F. Boumahdi, A. Madani, and I. Cheurfa, "Identifying depression in tweets using CNN-deep and BiLSTM with attention model," *International Journal of Web Applications*, vol. 12, no. 2, pp. 47–61, 2020.

BIOGRAPHIES OF AUTHORS



Rula Kamil    is a master's student at the Department of Software in Information Technology, University of Babylon, Iraq. As well as works as a programmer at University of Babylon since 2016. She obtained a bachelor's degree in Computer Science from the College of Science, University of Babylon in 2012. She can be contacted at email: rola.kamil@uobabylon.edu.iq.



Ayad R. Abbas    is a principal lecturer in Applied Computing at the Faculty of Computer Science and Environment at Technology University. He has extensive research interests covering a wide variety of interdisciplinary perspectives concerning the theory and practice of Applied Computing in Text Mining. He has published well over 40 peer reviewed scientific publications and 3 book chapters, in multidisciplinary research areas including: technology enhanced learning and applied artificial intelligence. He can be contacted at email: ayad.r.abbas@uotechnology.edu.iq.